Big Data Risk Analysis for Railway Safety

Coen van Gulijk, Peter Hughes, Miguel Figueres-Esteban, Valentina Viduto

Institute for Railway Research, University of Huddersfield, UK

1. Introduction

Computer scientists are quite clear in their belief that the internet is coming of age. They have a firm belief that the enormous amounts of data floating around in the internet will unchain a management revolution of uncanny proportions[1] [2]. This revolution is referred to as 'Big Data'. Yet, to date, the potential benefit of this revolution is scantily investigated for safety and risk management of the railways. This work reports about an investigation how Big Data can contribute to safety systems for the GB railways. The experience that is gained also sheds light on Big Data as a driver for change in the railway industry as a whole.

Cheap computer power supports the Big Data revolution. It allows for the use of large volumes of data in the railway industry. In the UK, Network Rail and ATOC are making substantial investments in their decision support tools for asset management[3] and accessibility of train running information[4]. Train running information is used in *thetrainline app* which has been downloaded 6.5 million times and directly adds value for travellers. Network rail's chief executive Mark Carne emphasized that safety remains a top priority for the GB railways in his George Bradshaw Address on February 25, 2015[5]. He emphasized the importance of the Digital Railway as one of the major efforts to drive the GB railways forward. The developments in this paper provide a glimpse of the Big Data phenomenon which will invariably be introduced in all facets of the railways.

RSSB and the University of Huddersfield have established a collaborative research programme to investigate the potential of Big Data techniques for safety in the railways in a research programme called Big Data Risk Analysis (BDRA). BDRA works on the intersection of the digital railway, (big) data-analytics, and railway safety. Some initial developments were made towards the analysis of large volumes of safety relevant data to support safer decision making and risk analysis. The preliminary results are promising but a broader discussion with industry and academia is called for.

[1] Mayer-Schönberger, V & Cukier K (2013) Big Data: a revolution that will transform the way we live, work an think. John Murray Ltd. London.
[2] McAfee, A. & Brynjolffson E. (2012) Big Data: the management revolution, Harvard Business Review Oct 2012: 61 – 67.
[3] See P40 in report 'Asset Management strategy'on http://www.networkrail.co.uk/aspx/12210.aspx?cd=3
[4] See: http://www.atoc.org/about-atoc/national-rail-enquiries/access-to-information/
[5] http://www.networkrail.co.uk/Mark-Carne-lifts-the-bonnet-on-Network-Rail-in-key-speech/

2. GB Railway industry vision

The GB Rail Technical Strategy Report 2012 (RTS)[6] states that the number of passengers on trains will continue to grow. It describes a vision where the GB railways contribute to the growth of the economy by ensuring customer satisfaction and value for money by being safe, reliable, resilient, meeting capacity and being service oriented. Six innovation themes were defined to support these objectives: control, command and communication; energy; infrastructure; rolling stock; information and customer experience. These themes heavily depend on data. To make an optimal use of data several organizations in the GB railways have made relevant rail-data available for a wider audience. Network Rail and ATOC are currently providing access to live data-streams about trains and tracks: BPLAN, Corpus, Movement, RTPPM, Schedule, SMART, TD, TSR, VSTP, Fares Data, Timetable Data and Routing Data[7] [8]. Very recently, ATOC have launched the 'data to improve customer experience' challenge[9] where universities and commercial companies can submit their solutions for serving the customer with software tools to improve their travel experience.

RSSB and the University of Huddersfield are working from a similar starting point as the GB RTS and ATOC. Having recognized the availability of data, the hypothesis is that Big Data techniques could be used to improve safety on the GB railways.

3. Current Big Data Risk Analysis projects

3.1 RAATS

RAATS is an acronym of Red Aspect Approach To Signals. The software accesses the TD live feed from Network Rail[7] to determine whether trains are approaching a signal at danger. Figure 1 shows the RAATS graphical interface. The pie-chart demonstrates that 23% of trains approach signal ET776 at danger between the 25th of March and the 13th of October 2014[10]. The software is flexible in the sense that either single signals, a group of signals or entire regions may be selected for different time frames. It can also differentiate between train types, e.g. local, intercity, freight.

This software may be used when the signaling system in the area has changed or to analyse whether this particular signal suffers from an exceptional number of red aspect approaches. The software helps to understand safety at red signals and can be used for better informed safety decisions when redesigning signaling systems. More information is given elsewhere[11].

[6] See: http://www.futurerailway.org/Documents/RTS%202012%20The%20Future%20Railway.pdf
[7] See: http://www.networkrail.co.uk/data-feeds/
[8] See: http://www.atoc.org/about-atoc/rail-settlement-plan/data-feeds/
[9] See: http://rruka.org.uk/events/data-to-improve-the-customer-experience/
[10] Stow, J, Zhao, Y. & Harrison, C. (submitted manuscript to IET) Estimating the Frequency of Trains Approaching Red Signals – a Key to Improved Understanding of SPAD Risk.
[11] Stow J, Zhao Y & Harrison C (2015) Estimating the frequency of trains approaching red signals – a key to improved understanding of SPAD risk, (submitted manuscript to Journal of Engineering).

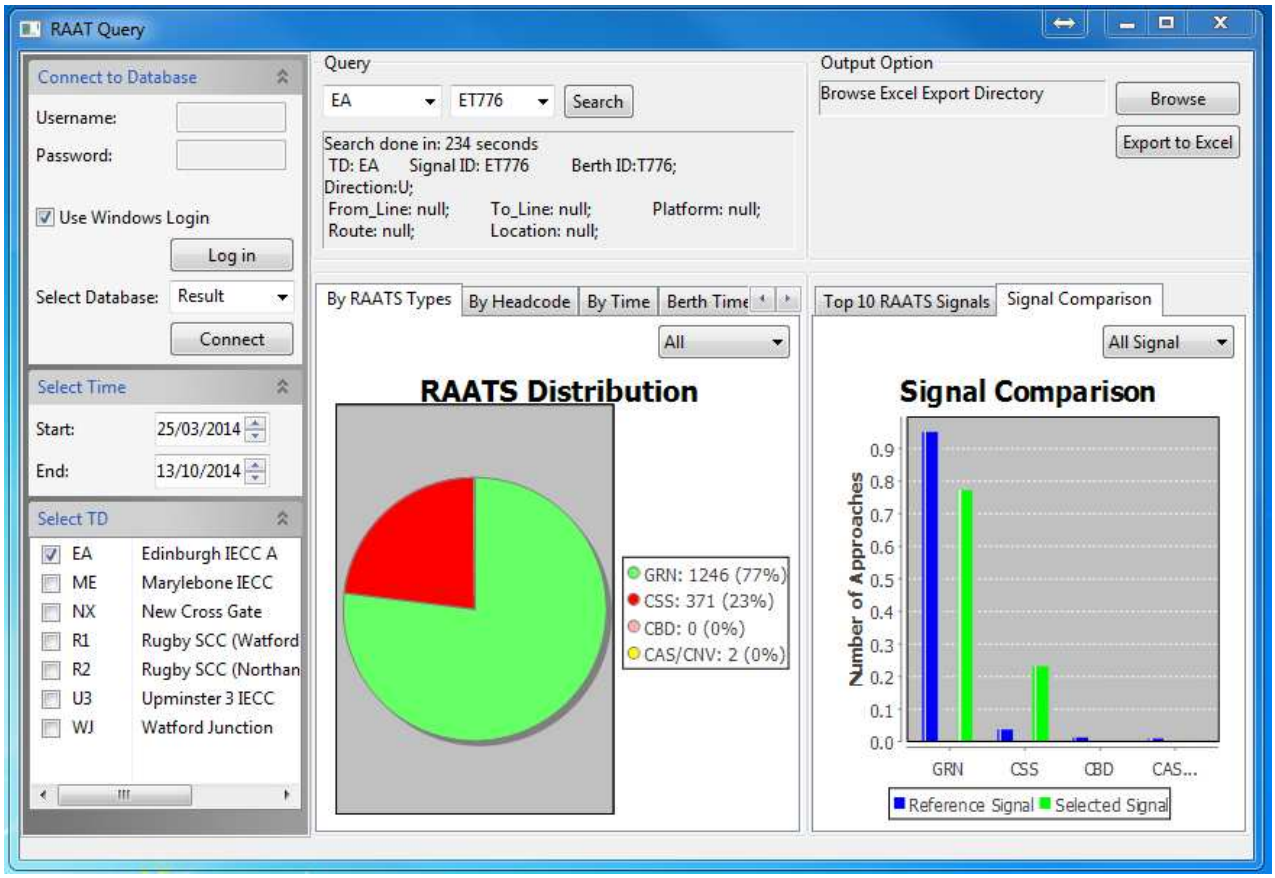

Figure 1: RAATS interface: signal ET776 for March-October 2014.

3.2 Learning from Close Calls

Network Rail have developed the Close Call reporting system[12] as a tool to gather safety-relevant information for the GB network that would not normally be reported as an incident in the SMIS database[13]. Close Call reports consist of a mixture of structured, categorized data and freeform text. The most important part is the freeform text; this is where employees report dangers in their own words. Due to the large number of freeform records (73,000 entries in the current dataset), it was impractical to manually review these records and therefore computer-based techniques were used. Computer-assisted analysis of freeform text is a complex task and it has received a great attention in the computer science field. Nevertheless, such techniques were not applied for the analysis of Close Call records before. The work to date has concentrated on the design of automated Natural Language Process software (NLP) where safety-relevant terms are identified and tagged with terms. These terms were derived from a list of about 660 safety-relevant concepts (and many more letter-permutations for those concepts) and 15,000 location names. Currently, these methods can be used to automatically extract Close Call records that relate to a certain incident. Figure 2 shows automatically extracted Close Calls related to trespass as function of time-of-day. A full description of the method is given elsewhere[14].

[12] See: https://www.safety.networkrail.co.uk/alerts-and-campaign/close-call
[13] See: http://www.rssb.co.uk/risk-analysis-and-safety-reporting/reporting-systems
[14] Hughes P & Figueres-Esteban M (2015) *Learning from close call events*, Report: IRR 110/89, IRR, Huddersfield. (Available in SPARK)

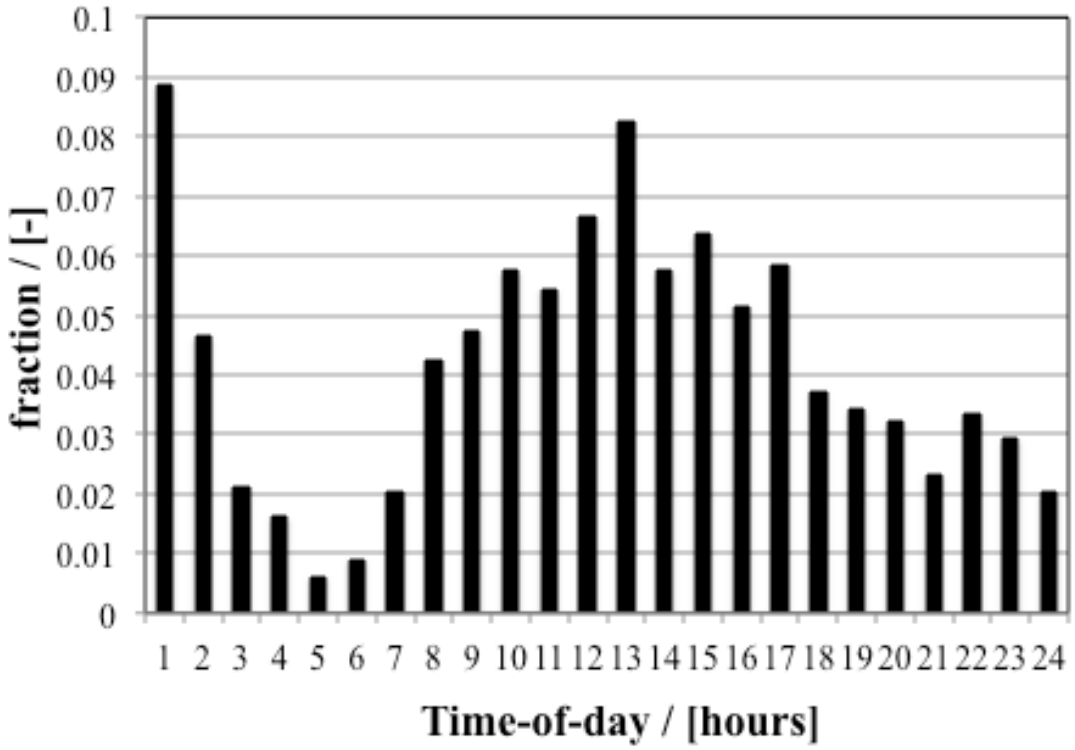

Figure 2: Close Call records as function of time-of-day as extracted with NLP methods.

4. Challenges for BDRA

The efforts for developing BDRA have started in September 2014. We believe that BDRA can be viable if it either produces a step-change in safety performance for the GB railways or if current safety performance levels can be maintained at significantly lower operation cost. Our initial efforts lead us to believe that BDRA could contribute to those aims. However, there are significant challenges to overcome. This section briefly outlines the challenges the BDRA team is currently dealing with.

4.1 Safety and risk

Though data analytics techniques developed by computer scientists around the world we found no clear link to safety sciences. The close call project has taught us that standard machine learning techniques do not yield satisfactory results for identifying meaningful safety data. Though we are investigating more advanced Machine Learning techniques today, we believe that safety and risk knowledge has to be codified into ontologies based on traditional risk tools for meaningful interpretation of data. We believe this to be key to the success of BDRA because there is little tolerance for errors when it comes to safety tools that are supposed to keep the railway system safe. We believe that, as a consequence, internet-based safety and risk tools will primarily be analysis techniques that inform decision makers rather than automated control systems.

Another important problem is the signal-detection problem; data analytics techniques are typically used for finding trends in large volumes of data. For safety purposes, the number of relevant signals might be very low and the quality of the data might be low. At this point it is hard to predict whether automated search engines could be developed to detect such signals or even whether meaningful causal correlations can be found using machine learning techniques.

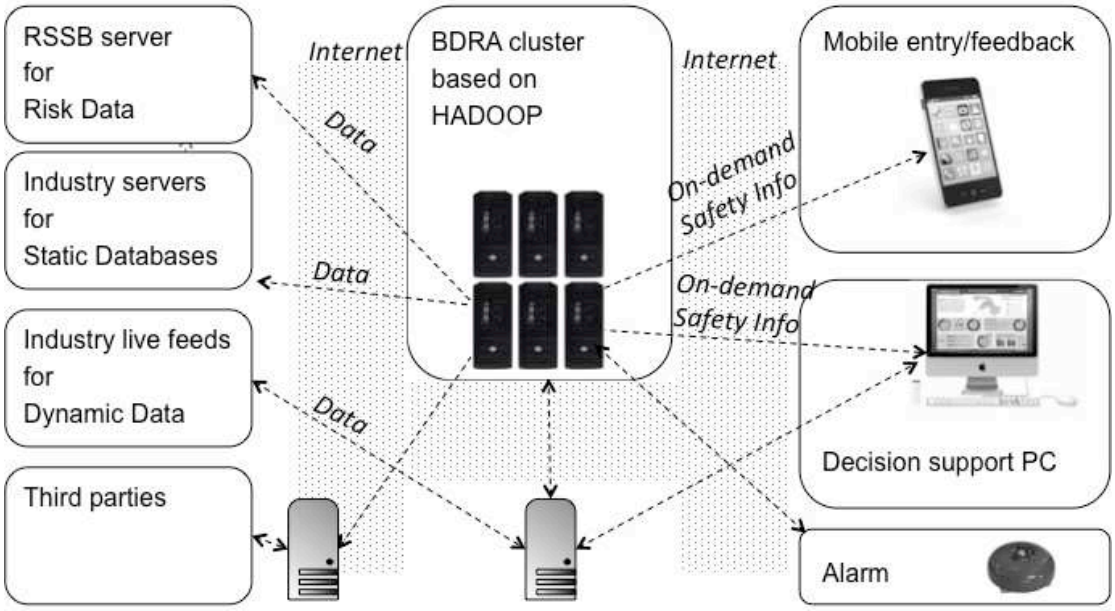

Figure 3: Technical system for BDRA.

4.2 BDRA Architecture Framework

The discussion above motivated a need for a new approach to the BDRA architecture that would address major challenges related to component technologies and data properties. Considering known best practices in defining architectures for new technologies, such as NIST Cloud Computing Reference Architecture, Intercloud Architecture Framework, and recent discussions by the NIST Big Data Working group, we have emphasised five components that address BDRA system:

- Data models, structures, types
- Data management, provenance, archiving
- Data analytics tools: BDRA software applications, visualisation, presentation
- BDRA infrastructure, including storage, computational power, network, operational support

It is likely that all data-feeds that are available today carry safety-relevant information and many more data-feeds might have to be monitored in the future. We expect that each data-feed would have a similar live-feed reading capability as RAATS has today so the software for reading the live feeds could grow. To support possible applications and services the following data centric applications have to be considered: hadoop related services and tools, cluster services, databases, NoSQL, parallel processing databases. Some of these tools are offered by the major cloud providers, such as Elastic Map Reduce, Dynamo, IBM Big Data Analytics, Cloudera, however at this development stage we need to understand the overall architectural requirements.

The top level BDRA system is shown in Figure 3. To address computational time issues we build our initial BDRA analytics on a Hadoop cluster as a scalable solution for future growing data demands. Data processing from various live feeds and other data sources requires data preservation that should allow data re-use, secondary search on the processed data and/or obtained results. However, this is only possible when cross-referencing, linkage and data identification are implemented.

4.3 Visualization

Although more than 300 research papers have been identified about this issue, the first results from that literature review process seem to indicate that the application of Information Visualisation (InfoVis) systems to Big Data is not clear and new challenges are emerging. Computer scientists invent new methods of the visualization of data yielding literally hundreds of different visualization techniques. There does not seem to be a clear shared view that informs us about the right way to represent risk data from BDRA. Apart from that, visualization techniques are not just used for representation of results, they can also play a part in selection of data-sets, representation of ontologies and software-monitoring tools. But the problem becomes more challenging when visualisation techniques have to be tuned to the audience of BDRA. Visualisation, despite being largely discussed among computer scientists, also involves in-depth understanding of psychology and risk.

4.4 Summary

The experience gained with these projects suggests that there is added value to the Big Data approach for Risk analysis but that much work remains. This paragraph describes the first steps toward an integrated railway safety system based on Big Data analysis. It is clear that there is a long way to go before an integrated system will be ready but these first steps show promise.

5. The bigger picture

The development of BDRA is set against the Big Data revolution. We believe that it is likely that railways around the world will feel its effect. Mayer-Schönberger & Cukier[1] take the fact that data is rapidly amassing as evidence for this. The railway industry, and the world at large is rapidly increasing its data-position with data-intensive systems such as SCADA, condition monitoring, GPS locations, time tables, payment data, social media and so forth. Simply the fact that there is so much data out there begs for the development of tools to harness value from that. Today, the development of sensible analysis tools is no longer the prerogative of Google, Facebook and IBM alone. A whole new industry is developing where smart analysis tools are being developed. The GB railways have encountered this industry and harnessed value through *thetrainline app* but that is just the beginning.

Some of the concepts in Big Data are controversial. For instance, for Big Data applications to work, the emphasis on correlations that are found in data rather than causal relations. That is to say, the emphasis is mostly on finding correlations without trying to find out why these correlations exist. For big data analysts the correlation is ‘usually good enough’ but this is a leap of trust for the railway industry. Also, the data (and sometimes the correlations) are messy. This is mostly due to the fact that a number of data-sources are combined. Often, one or more databases are not well structured or incorporate social media where misspelling, mistakes and outright nonsense can obfuscate valuable correlations. With the Big Data approach amassing data till the point of saturation is the cure, even weak correlations can be spotted if the data-set is large enough. Also, railway industry partners amass lots of data but are not necessarily willing to share that information. Fortunately, that does not necessarily have to be the case for Big Data to be useful. If the owner of the data can provide key figures based on in-house data-analytics there is no need to just give away valuable data.

Yet, the railways provide the right soil for Big Data projects. It is an equipment-rich industry for which more and more monitoring options open up. Operations information is constantly passed back and forth to ensure smooth operation. Trains provide service to millions of travellers with

access to smart-phones and all the Big Data boons associated with them and, last but not least, the economy of railways is key to it success. If the railways can overcome the controversies, data utilization may well be the key to improved performance in the railway industry from now right up to 2050.

6. Conclusion

This paper gives a brief overview over Big Data Risk Analysis for the GB railways. GBRA is a new development for safety and risk analysis based on modern data-analytics and huge amounts of data that are generated by the GB railways. The novelty mostly lies in the use of software tools that are not normally used in the railway industry nor for safety and risk. The examples presented here show that initial efforts show promise for the future even if the research project has just started. But this work points to a benefit beyond BDRA for the GB railways: the combination of railway engineering, an emphasis on information, contact with the public and rigorous financial boundaries, the railways combine all the ingredients in which data-analytics applications have been successful before. We believe that the development of sensible data-analytics will help the railways to improve their performance from now right up to 2050 and that BDRA contributes to that.

7. Acknowledgements

RSSB is gratefully acknowledged for their funding under the UoH-RSSB Strategic partnership for Rail Safety Modelling.